DOI:10.1478/C1A0601006

Atti dell'Accademia Peloritana dei Pericolanti
Classe di Scienze Fisiche, Matematiche e Naturali
Vol. LXXXIV, C1A0601006 (2006)
Adunanza del 28 novembre 2005

VOCAL TECNOLOGY: A NORMALIZATION APPROACH

ALFIO PUGLISI

(Nota presentata dal Socio Emerito Maria Teresa Calapso)

ABSTRACT. From the 1990s onwards the use of digital technology for voice and image transmission (GSM mobile telephones, satellite transmissions and Frame Relay and ATM networks) has brought about the convergence of information technology and telecommunications, leading to the birth of the ICT (Information & Communication Technologies) sector. Currently, internal telephone networks, LANs, internet connections and geographical data transmission networks are being unified in most organizations of a certain size.

1. ASR (Automatic Speech Recognition)

The first step in the recognition of spoken language consists of sampling the voice, filtering it so as to attenuate any background noises. The next step consists of identifying the phonemes, an operation hindered in some cases by the peculiar characteristics of the speaker, such as tone of voice or accent. Finally, after having understood the phonemes, the program arranges them into morphemes and words.

ASR systems rely on a grammar which defines a set of valid expressions through which the user can interact with a vocal application. We present a possible classification of Speech Recognition Grammars proposed by sun Microsystems:

Rule-Based Grammars. The recognition process is bound by a series of rules, created in such a way so as to give the user a certain amount of freedom of expression, while limiting the vocabulary available in order to make recognition as fast and accurate as possible.

Dictation Grammars. These systems impose fewer restrictions than previous ones, but consequently require greater computational power and make more recognition errors; in some cases, systems "with dictation" ask the user to pause between one word and another.

Let us look at some of the languages used for the definition of speech recognition grammars:

Nuance Grammar Specification Language. GSL is a proprietary standard developed by Nuance; it allows the developer to plan sophisticated grammars, both of the static and dynamic kind, and it includes an optimization system to speed up the runtime.

Speech Grammar Markup Language. GRXML is a an XML-based language developed by the members of W3C; it is currently under development and only allows the construction of grammars with a very simple set of commands.

Java Speech Grammar Format. JSGF is a symbolism independent of the platform and implementation, consistent with the JAVA language; it defines a rule-based grammar.

Some of the ASR motors on the market are VoxGatewayServer By Motorola, which uses Nuance automatic speech recognition technology, and IBM Websphere Voice Server, a family of servers providing vocal access to the Web.

1.1. ASR-Dependent Speaker Identification. We will make use of the following notation when describing the ASR-dependent speaker identification approach and its corresponding normalization methods: Let X represent the set of feature vectors, $\{x1, ..., xN\}$, extracted from a particular spoken utterance. Let the reference speaker of an utterance X be $S(X)$. Furthermore, assume that the aligned phonetic transcription of the utterance, $\Phi(X)$, provides a mapping between each feature vector xk and it's underlying phonetic unit $\phi(xk)$. In our ASR-dependent approach, each speaker S is represented by a set of models, $p(x|S, \phi)$, which mirror the CD acoustic models trained for speech recognition, $p(x|\phi)$, where ϕ ranges over the inventory of phonetic units used in the speech recognizer. The use of a set of context-dependent phonetic models for each speaker is markedly different from global GMM modeling approaches, where the goal is to represent a speaker with a single model, $p(x|S)$. During evaluation, automatic speech recognition is performed on the utterance producing an automatically generated phonetic transcription, , which assigns each vector,xk, to its most likely phonetic unit, . The phone assignments generated during speech recognition can then be used to calculate speaker-dependent phone-dependent conditional probabilities, $p(x|S, p(x\,|S\ \widehat{\phi}(x))$. Ideally, these probabilities alone would act as suitable speaker scores for making a speaker identification decision. For example, the closed-set speaker identification result might be:

$$\widehat{S} = (X) = \arg\ \underset{s}{Max}\ p(X\,|S, \Phi, X)) \quad . \tag{1}$$

In practice however, enrollment data sets for each speaker are typically not large enough to accurately determine the parameters of $p(x|S, \phi(x))$ for all $\phi(x)$.

2. Speaker Adaptive (SA) Normalization

We originally described a speaker adaptive normalization approach in [4]. This technique relies on interpolating speaker dependent (SD) probabilities with speaker independent (SI) probabilities on a per-unit basis. This approach learns the characteristics of a phone for a given speaker when sufficient enrollment data is available, but relies more on general speaker independent models in instances of sparse enrollment data. Mathematically, the speaker score can be written as:

$$Y(X, S) = \frac{1}{|X|} \sum_{x \in X} \log \left[\lambda_{S,\Phi(x)} \frac{p(x\,|S\ \Phi(x))}{p(x\,|\Phi\ (x))} + \left(1 - \lambda_{S,\Phi(x)} \frac{p(x\,|\Phi\ (x))}{p(x\,|\Phi\ (x))} \right) \right] \quad . \tag{2}$$

Here, $\lambda_{S,\Phi(x)}$, is the interpolation factor given by:

$$\lambda_{S,\Phi(x)} = \frac{n_{S,\Phi(x)}}{n_{S,\Phi(x)} + \tau} \ . \tag{3}$$

In this equation, $n_{S,\Phi(x)}$, refers to the number of times the CD phonetic event $\widehat{\varphi}(x)$ was observed in the enrollment data for speaker S, and τ is an empirically determined tuning parameter that was the same across all speakers and phones. By using the SI models in

the denominator of the terms in Equation 2, the SI model set acts as the normalizing background model typically used in speaker verification approaches. The interpolation between SD and SI models allows our technique to capture detailed phonetic-level characteristics when a sufficient number of training tokens are available from a speaker, while falling back onto the SI model when the number of training tokens is sparse. In other words, the system backs off towards a neutral score of zero when a particular CD phonetic model has little or no enrollment data from a speaker. If an enrolled speaker contributes more enrollment data, the variance of the normalized scores increases and the scores become more reflective of how well (or poorly) a test utterance matches the characteristics of that speaker's model.

3. Phone Adaptive (PA) Normalization

An alternative and equally valid technique for constructing speaker scores is to combine phone dependent and phone independent speaker model probabilities. In this scenario, the speaker-dependent phone-dependent models can be interpolated with a speaker-dependent phone-independent model (i.e., a global GMM) for that speaker. Analytically, the speaker score can be described as:

$$Y(X,S) = \frac{1}{|X|}\sum_{x\in X}\log\left[\lambda_{S,\Phi(x)}\frac{p(x\,|S\ \Phi(x))}{p(x\,|\Phi\ (x))} + \left(1-\lambda_{S,\Phi(x)}\frac{p(x\,|S\,)}{p(x)}\right)\right] \tag{4}$$

Here, $\lambda_{S,\Phi(x)}$ has the same interpretation as before. The rationale behind this approach is to bias the speaker score towards the global speaker model when little phone-specific enrollment data is available. In the limiting case, this approach falls back to scoring with a global GMM model when the system encounters phonetic units that have not been observed in the speaker's enrollment data. This is intuitively more satisfying than the speaker adaptive approach, which backs off directly to the neutral score of zero when a phonetic event is unseen in the enrollment data.

4. Remarks and conclusions

For our experiments we have examined both the closed-set speaker identification and speaker verification problems. Because our data is collected via individual calls to our system, we can evaluate speaker identification at both the utterance level and the call-level. In our case the utterance-level evaluation could be quite challenging because any single utterance could be quite short (such as the single word utterance "no") or ill-matched to the caller's models (as might be the case if the caller uttered a new city name not observed in his/her enrollment data).

In many applications, it is acceptable to delay the decision on the speaker's identity for as long as possible in order to collect additional evaluation data. For example, when booking a flight, the system could continue to collect data while the caller is browsing for flights, and delay the decision on the speaker's identity until the caller requests a transaction requiring security, such as billing a reservation to a credit card. To simulate this style of speaker identification, we can evaluate the system using all available utterances from each call in the evaluation data.

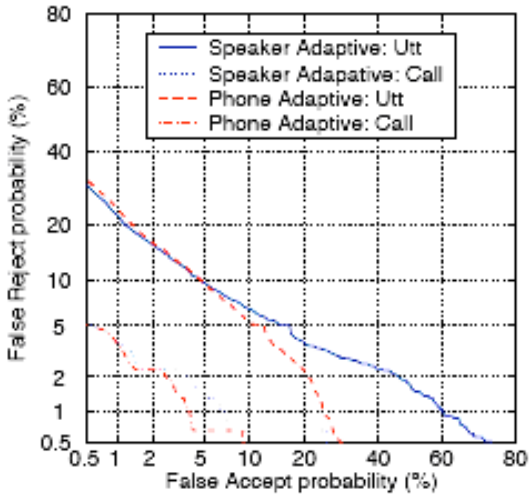

FIGURE 1. DET curves showing false rejection probability versus false accept probability for speaker adaptive vs. phone adaptive normalization.

4.1. Comparison of normalization schemes. We performed several experiments. First, we compared the performances of the two normalization approaches on the task of closed-set speaker identification using the 3705 in-set utterances. The identification error rates are shown in Table 1. We see that using the full amount of enrollment data per speaker, both techniques perform equally well. On limited enrollment data per speaker, the phone-adaptive normalization approach performs better. This is presumably because it retains the ability to discriminate between speakers even when there are many instances of sparsely trained phonetic units.

For the speaker verification task, we used the 2946 out of set utterances to perform same-gender imposter trials (i.e., each utterance was only used as an imposter trial against reference speakers of the same gender). The detection error trade-off (DET) curve of the two approaches is shown in Figure 1. For all four curves, the models were trained on all available data. From the region of low false acceptances through the equal error rate

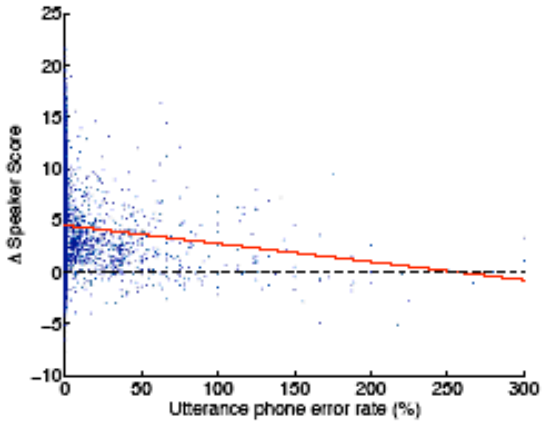

FIGURE 2. Plot of correct speaker discrimination versus utterance phone error rate using MT Models. Each point represents a single utterance. The horizontal axis shows the phone recognition error on the utterance. The vertical axis indicates the difference between the scores of the reference speaker and the best competitor (negative values indicate identification errors). A best-fit linear approximation of the data is superimposed.

region of the DET curve, the two normalization techniques have very similar performances. However, in the "permissive" operating point region with low false rejection rates, the phone-adaptive approach has significantly lower false acceptance rates than the speaker adaptive approach. This observation is important for a conversational dialogue system where convenience to frequent users is a factor. For example, if we want to maximize convenience by ensuring that only 1% of true speakers are falsely rejected, then the speaker adaptive method will result in a 60.3% false acceptance rate, while the phone adaptive method will result in a 24.6% false acceptance rate.

Amount of Enrollment Data	Speaker ID Error Rate(%)	
	SA Norm	PA Norm
Max 30 utts	26.4	22.5
Max 100 utts	18.4	15.9
All available	9.6	9.6

TABLE 1. Closed set speaker identification error rates on individual utterances for different amount of enrollment data speaker for adaptive vs. phone adaptive normalization.

In this paper, we addressed the issues of speaker score normalization and of using automatically generated transcriptions for training speaker models when performing ASR-dependent speaker identification.

We found that using a phone-adaptive approach is beneficial for normalizing speaker scores compared to a speaker-adaptive approach. Although both methods have similar speaker identification performance, the phone-adaptive method generates scores that are more stable on speaker verification tasks, yielding fewer false acceptances of imposters at permissive operating points where low false rejection of known users is desirable. In comparing the models trained from manually and automatically generated transcriptions, we found no significant differences in speaker discriminability between the two approaches. This discovery indicates that we can take an unsupervised approach to training speaker models without adversely affecting our speaker identification results.

References

[1] U. Chaudhari, J. Navratil, and S. Maes, "Multi-grained modeling with pattern specific maximum likelihood transformations for text-independent speaker recognition", IEEE Trans. Sp. Aud. Proc. **11**(2), 100 (2003)

[2] J. Glass, "A probabilistic framework for segment-based speech recognition", Computer Speech and Language **17**(2-3), 137 (2003)

[3] T. J. Hazen, D. A. Jones, A. Park, L. C. Kukolich, D. A. Reynolds, "Integration of speaker recognition into conversational spoken dialogue systems", in *Proc. EUROSPEECH*, Geneva, Switzerland, September 2003, pp. 1961-1964

[4] A. Park and T. J. Hazen, "ASR dependent techniques for speaker identification", in *Proc. ICSLP*, Denver, Colorado, September 2002, pp. 2521-2524

[5] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", Speech Communications **17**(1-2), 91 (1995)

[6] S. Seneff, C. Chuu, and D. S. Cyphers, "Orion: From online interaction to off-line delegation", in *Proc. ICSLP*, Beijing, China, October 2000, pp. 767-770

[7] S. Seneff and J. Polifroni, "Formal and natural language generation in the Mercury conversational system",in *Satellite Dialogue Workshop of the ANLP-NAACL Meeting*, Seattle, WA, April 2000

[8] F. Weber, B. Peskin, M. Newman, C. Andres, L. Gillick, "Speaker recognition on single- and multispeaker data", Digital Signal Processing **10**(1), 75 (2000)

Alfio Puglisi
University of Messina
Faculty of Economy
Via dei Verdi 75, 98122 Messina, Italy
E-mail: puglisia@unime.it

Presented: November 28, 2005
Published on line on November 3, 2006